The 11th International Scientific Conference
eLearning and Software for Education
Bucharest, April 23-24, 2015
10.12753/2066-026X-15-000

A SEMANTIC APPROACH TO ANALYZE SCIENTIFIC PAPER ABSTRACTS

Ionut Cristian Paraschiv, Mihai Dascalu, Stefan Trausan-Matu
Computer Science Department, University Politehnica of Bucharest, 313 Splaiul Independentei, Bucharest Romania
ionut.paraschiv@cti.pub.ro, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro

Philippe Dessus
LSE, Univ. Grenoble Alpes, 38040 Grenoble CEDEX 9 France
philippe.dessus@upmf-grenoble.fr

***Abstract**: Each domain and its underlying communities evolve in time and each period is centered on specific topics that emerge from textual sources that characterize the domain. Our analysis represents an extension of other researches performed on the same corpora that were focusing more on evaluating co-citations between the articles in order to compute their importance score (Grauwin and Jensen [1]). Our approach presents a general perspective of the domain by performing semantic comparisons between article abstracts using natural language processing techniques such as Latent Semantic Analysis, Latent Dirichlet Allocation or semantic distances in lexicalized ontologies, i.e. WordNet. Moreover, graph visual representations are generated using Gephi in order to highlight the keywords of each paper and of the domain, the document similarity view or the table of keyword-abstract overlap score. The purpose of the views is to minimize the learning curve of the domain and to facilitate the research process for someone interested in a particular subject. Also, in order to further argue the benefits of our approach, some potential refinements of the methods for classification that can be performed as future improvements are presented.*

***Keywords:** Scientometrics; discourse analysis; semantic similarity; extraction of domain key concepts*

I. INTRODUCTION

Every researcher has to deal every day with the problem of annotating and finding papers from a specific domain. Moreover, one must be able to perform a deep search for related articles, as it is very important to study the current results obtained in the field in order to be able to deliver something new. In this context, the necessity of building a paper comparison tool is very important, as it can decrease the time for finding important information regarding a given topic and it can give a general representation of the domain, being useful not only for researchers, but also for people interested in finding more information about a specific subject.

In this paper we propose a semantic analysis of the article abstracts extracted from the citation index Web of Science, from the category Education and Educational Research, taken between the years 2000-2004. This study begins with co-citation analysis [2], an existing method for paper comparison using co-citations, continues with details about our system, along with the methods used as a background, and finishes with future work and conclusions.

1.1 Co-citation Analysis

Co-citations analysis is frequently used to build an article graph of the domain [2]. The underlying idea is that the semantic meaning of papers can be extracted by analyzing their place in the graph and the connections with their neighbors. Paper *A* is connected to paper *B* if both cite at least one common reference (co-citation). In this particular situation, a directed link from *A* to *B* is added to

the article graph, and its weight is related to the number of co-citations. Therefore, by using this method, the most central and important documents can be determined from the domain. This approach is co-occurrence-based and bibliography-centered, thus fast to process with very large sets of papers. Moreover, bibliographic coupling captures collaboration and communities, making the extraction of the article network easier. However, citation strategies are influenced by researchers' practices [3] and motivations [4]. Usually, the citation number differs across domains and is influenced by authors' impact factor increasing endeavors. Another disadvantage is that the co-reference graph does not address the problem of multiple co-references of the same publication inside a paper [4]. Also, two citations of a given paper may refer to different aspects of the cited paper, and therefore loosing semantic meaning.

1.2 Semantic Comparison of Papers

The approach used is our system consists in computing article similarity based on their abstracts, which usually contain the central aspects of a paper [5]. This content-based approach can be accurate for measuring cohesion or topic similarity. Recent research [4] has proved that LDA (Latent Dirichlet Allocation [7]) adequately captures topics used in large corpora and can be used for text analysis and summarization. The downside is that semantic analysis is more computationally demanding in contrast to network extraction that is more difficult due to the fast corresponding expansion. The text from the abstract is analyzed using a pipeline described in the next sub-section, and the information is annotated using document comparison models like Latent Semantic Analysis (LSA) [6] and Latent Dirichlet Allocation (LDA). Usually this processing is highly costly from a computational point of view as it requires intensive hardware resources and the results are used to create different views of the domain using articles or important word graphs.

Compared to the co-citation model, the semantic comparison method can give some bad results as an author can co-quote an important document and include a lot of copied text, and his paper becomes automatically a central document within the graph. A workaround for this problem will be presented in the conclusions section.

1.3 Semantic Measures used within our Experiments

The experiment consists in using different models of semantic analysis for paper summaries to compare them directly, instead of using references as a proxy. Latent Semantic Analysis [6] is a natural language processing method that extracts the most important terms/concepts from a document as to compare them to other documents, in order to compute a similarity between them. The model uses a sparse term-document matrix on which a Singular Value Decomposition (SVD) is applied in order to project the concepts in a vector space with reduced dimensionality. To compare two documents from the initial set, a simple cosine similarity can be computed between their corresponding vectors from the SVD.

Latent Dirichlet Allocation [7] is a large corpora comparison model that uses a topic model. The model extracts topics from co-occurrence word patterns from a training corpus and provides similarity scores between documents based on their corresponding topic distributions.

WordNet [8] is one of the largest and most frequently used lexical ontology in English with more than 150.000 concepts. Using an ontology such as this one, one can compute semantic measures for estimating the relatedness between two words using the various relations from the ontology [9]. The similarity for two words can be computed for example using their attributes or the relations they have. The similarity can also be computed by going to a certain depth in the graph and examine the nodes. One of the best approaches is to combine multiple similarity measures, but unfortunately a tradeoff between accuracy and computational speed must be made.

For computing the similarity between two documents, the best approach would be to combine the previous three methods in order to compute an aggregated cohesion score [10]. An approach that can decrease the preprocessing time would be to use in the document-word matrix only the important terms from the initial set, that can be combined into topics. Usually, to determine the most important words, stop words must be removed (e.g., "*as*", "*the*", "*in*", "*a*", "*an*", etc.) and the first *n* words ordered by their appearance count must be extracted (*n* is a tradeoff between computational effort and accuracy). Also, these words must be weighted by their semantic relatedness to the whole meaning of the abstract. At the end, similar words can be combined inside topics. Before applying complex

models to the input text like the ones described earlier, it needs to be added in a Natural Language Processing (NLP) pipeline [11] in order to remove unnecessary data. The first step consists of spellchecking and stop words removal used for keeping only relevant and correct words that can be found inside a dictionary. After tokenization and splitting, the text is inserted into computational vectors, which are reduced to their morphological unit using stemming (i.e., Snowball stemmer [12]). Finding named entities inside the text is also an important process that provides additional semantic meaning and increases the overall accuracy. Additional benefits are obtained after performing co-reference resolution [13] as co-reference links indicate intra-textual cohesiveness.

II. SPECIFIC VIEWS HIGHLIGHTING DIFFERENT SEMANTIC SIMILARITY FACETS OF PAPER ANALYSIS

The goal of our model is to create a good representation of the papers, as to better understand the domain and therefore help researchers in their work. We used a corpus of 1,000 abstracts from Jensen & Grauwin [1] database, extracted from the educational science domain. We modified and extended *ReaderBench* [10; 14] in order to perform the current analyses. *ReaderBench* is a research software system used to evaluate documents from a cohesion-based perspective. It also supports chat analysis as to view the inter-animation of voices and to compute the impact of participants within the conversation [15], and it represented a great starting point for analyzing paper abstracts as it has already implemented many of the methods described in the previous chapter.

2.1 Document Similarity View

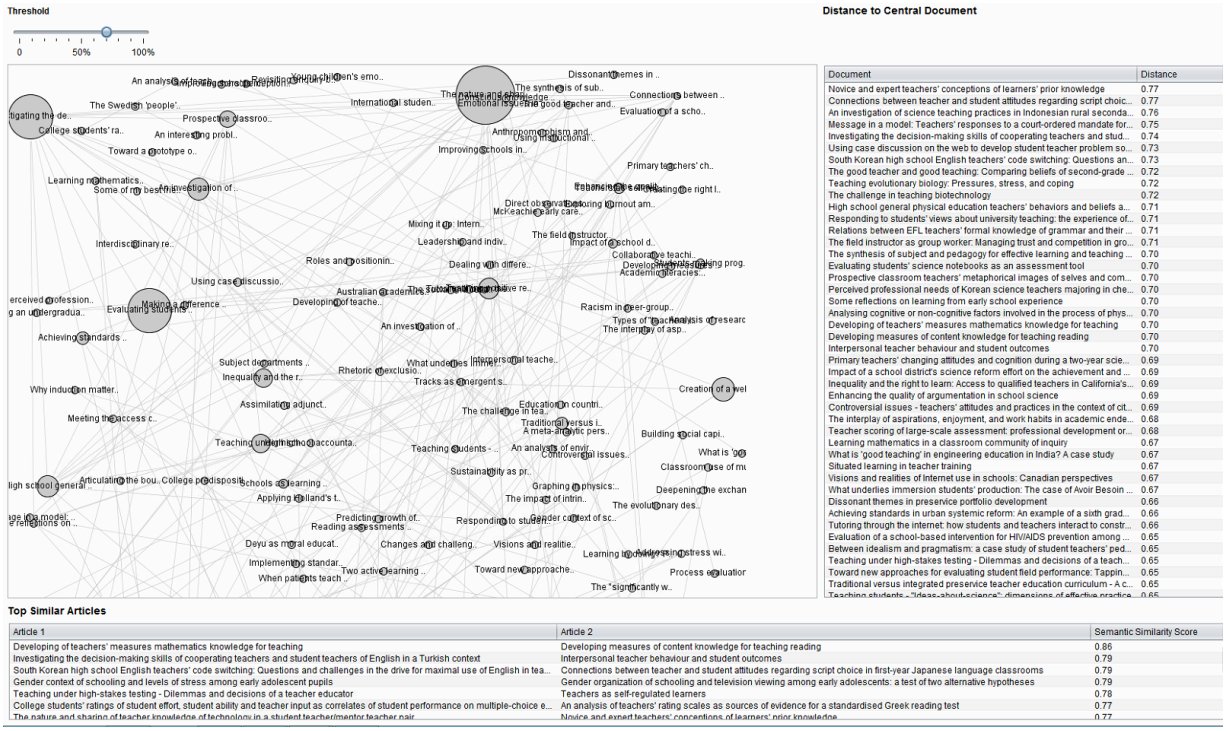

Figure 1. Corpus Similarity View

The first approach of document representation is to create a graph with all the documents. Two documents are connected if their similarity exceeds a certain threshold that can be selected by the user. Also, it can be interesting to check which are the most central documents, and therefore all the nodes from the graph have a size directly proportional on their centrality. The corpus similarity view also contains a table with the most similar documents, and another one with all the documents ordered by their distance to the central document. In the bottom of Figure 1, the top similar articles can be seen, a table that can be very useful if the central documents are very similar. In the right hand side of the figure, all the documents are ordered by their distance to the most central document, represented as the biggest node from the graph.

An important aspect to note is that, if a new paper is generated and it resembles the most central document from the graph, it can easily become one of the central nodes. This can be a serious problem, as anyone can create a similar paper to an important one from the graph, and the model would conclude that it is a central article from the document set. This problem can be overcome by combining with the results from the co-citation model.

2.2 Document Centrality View

Another approach a researcher could take when studying a certain domain is that, starting from a paper, to try to find other similar documents. The document centrality view (see Figure 2) first needs a central document *D* that can be selected by a user, and then displays all the most similar articles to *D*. The threshold is user-selectable, along with the depth level. For depth level 1, document *D* and all of its most similar documents *Ds* can be checked inside the graph. For a depth level 2, all the documents that are similar to *Ds* from step 1 are added in the graph. The depth level was limited to 3 as the search space grows exponentially and can exceed the computational level.

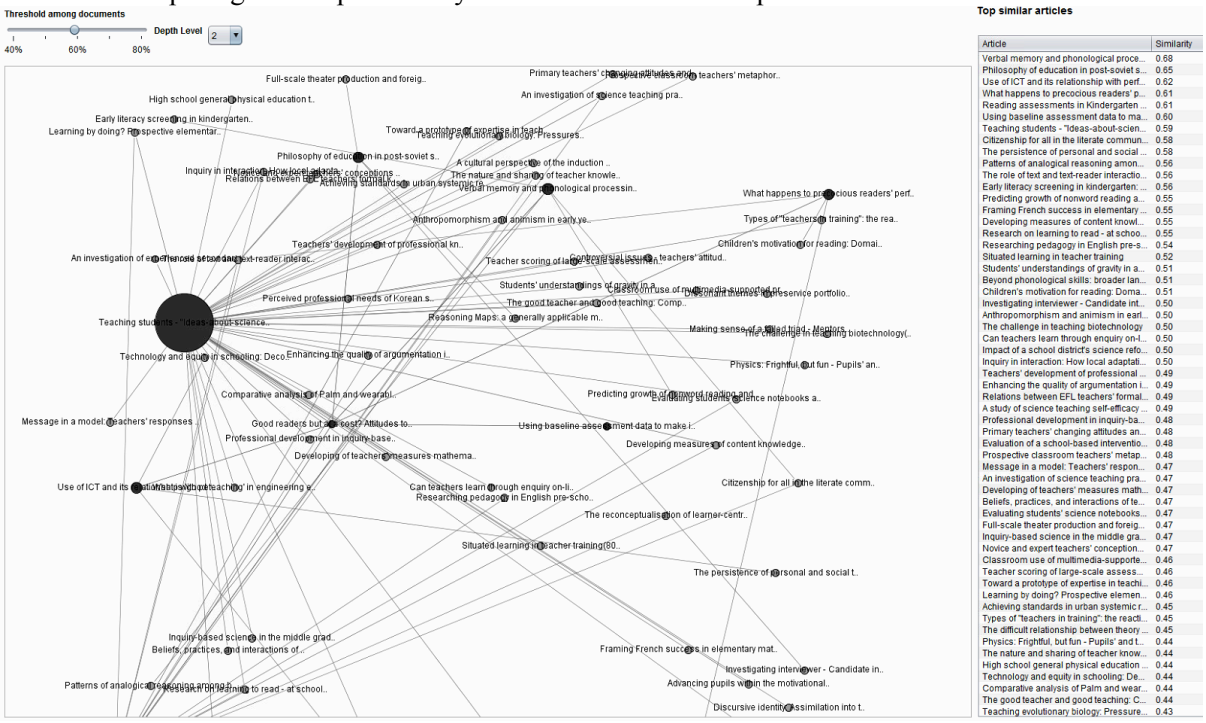

Figure 2. Document Centrality View

This view can be very useful when someone is interested in a given subject and needs to read more articles that are related one to each other. As the abstract usually contains the most relevant information about the specific subject, the best way to find other articles would be to extract the most similar papers from the domain set. By imposing a threshold, the user is responsible of the similarity level between the displayed articles. In order facilitate the visualization, all articles are displayed inside a table, ordered by their similarity to the initially selected document.

2.3 Concept View

Another interesting approach when studying a certain domain is to find the most important concepts from the field. With this in mind, the concept view creates a graph with the most important words from the document set, where the nodes (words) are sized according to their centrality. The words sorted by their centrality can be checked inside the table from Figure 3.

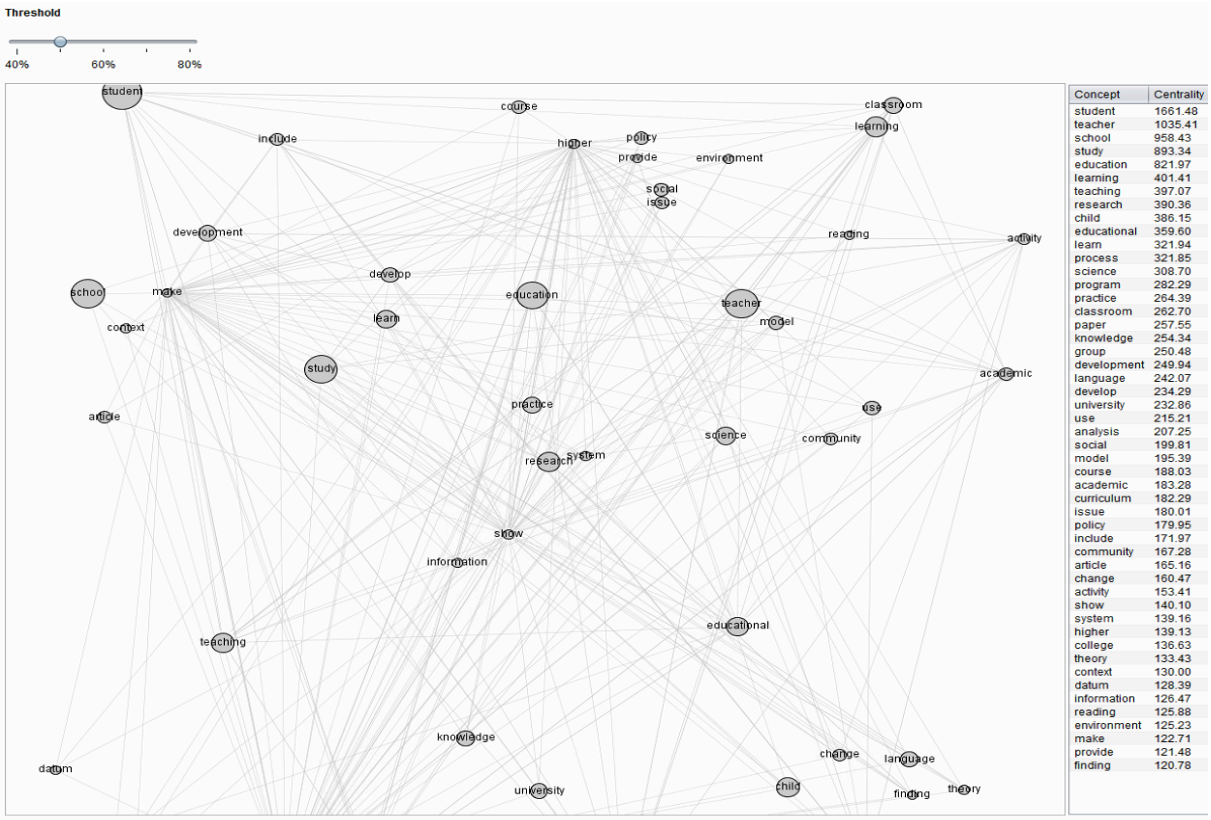

Figure 3. Concept View

As it can be easily observed from Figure 3, the most central words for the category "Education and Educational Research" are "*student*", "*teacher*", "*school*", "*education*", "*learning*", "*teaching*" etc., representative for the field. This approach confirms that the proposed semantic analysis can lead to an adequate representation and understanding of a domain and can generate a general perspective of the withheld knowledge.

2.4 Keyword-abstract Overlap

In addition to the previous view and in order to study how relevant the keywords assigned by the paper's author(s) are, when compared to the abstract, the keyword-abstract overlap view (see Figure 4) was created. Each document received a compound score composed from occurrences of keywords inside the abstract (30%) and the semantic relatedness between all the keywords and the abstract (70%). The previous weights were empirically assigned to best reflect the overall similarity by considering both lexical and semantic relatedness.

The results obtained were also encouraging as the displayed similarity scores reflect a reliable estimation of the coverage of the keywords in terms of the paper's abstract. For example, in Figure 4, the article from the first place has an aggregated score of 0.75 as all of its keywords are inside the abstract, are semantically related to it, and the abstract is detailed and comprehensive. Usually the articles from the last places have a small and evasive abstract with a limited number of keywords (one, two, maybe three) not retrievable within the textual description. This score can be used to compute the trustworthiness of an article and can be used in generating the views from subsections 2.1 and 2.2. For example, if the central paper from the document similarity view has a small abstract-keyword score, then it may be vague and its score can be correspondingly weighted.

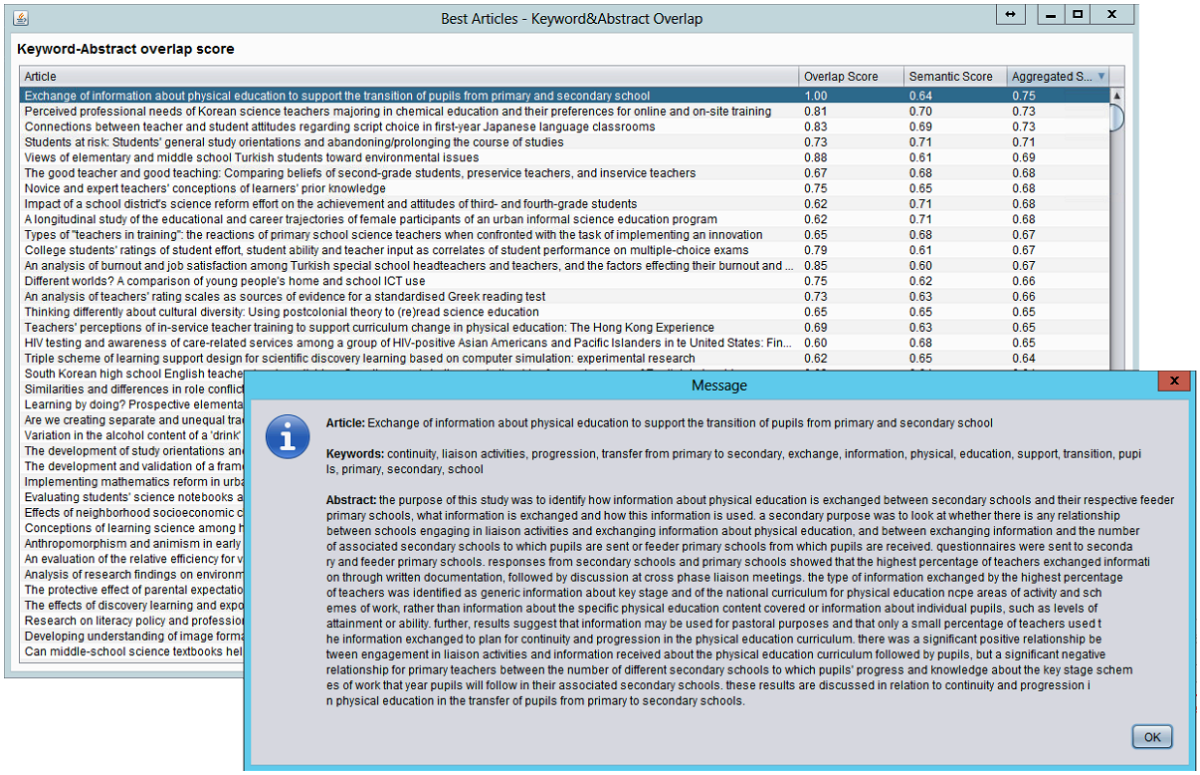

Figure 4. Overlap between article keywords and mentioned concepts

III. CONCLUSIONS AND FUTURE WORK

In this paper, some general methods of viewing the data from an initial document set using full text analysis were proposed, methods that can be further extended. Date-related analysis will allow to perform trends analysis, i.e., which keywords are mostly relevant in a given period of time. Moreover, sentiment analysis [16] will allow to check the authors' citation motivation using or to display the overall paper's semantic flow using inter-paragraph cohesion [17]. Such analyses are more complex than the one described in this paper, but it can help us increase the overall model's functions. Some other interesting future work that can be made based on the current model would be to focus on specific sub-fields, in order to build a domain's representation or semantic map. Also, the co-citation approach can be easily coupled with the abstract analysis method as to increase the overall quality of the analysis.

Moreover, the current model can be applied on a time frame as to check the effect of the introduction of bulks of published papers on the paper's space, as to check the dynamic impact and how the domain changes over the years. Even further, the model can be integrated in recommender systems (which papers are most suitable to read?), writing advices, or adding RDF paper metadata in order to facilitate semantic queries. All in all, the semantic analysis of paper abstracts is a good start for annotating papers and increasing the general representation and understanding of a certain domain.

Acknowledgements

The work presented in this paper was partially funded by the Sectorial Operational Programme Human Resources Development 2007-2013 of the Ministry of European Funds through the Financial Agreement POSDRU/159/1.5/S/134398.

We also thank Pablo Jensen and Sebastian Grauwin for providing the initial corpus of paper abstracts.

References

[1] Grauwin, S., & Jensen, P. (2014). Project EducMap. from Lyon University http://ife.ens-lyon.fr/ife/recherche/groupes-de-travail/educmap/educmap

[2] Boyack, K.W., & Klavans, R., 2010. Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately? *Journal of the American Society for Information Science and Technology. 61*(12). Pag 2389–2404

[3] Sword, H., 2012. *Stylish Academic Writing*, Harvard University Press. Cambridge, Massachusetts & London, England

[4] Song, M., & Ding, Y. , 2014. Topic modelling: Measuring scholarly impact through the topical lens In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring scholarly impact: Methods and practice*, Springer. Pag 346

[5] Ding, Y., Song, M., Wang, X., Zhang, G., C., Zhai, & Chambers, T. , 2014. Content-based citation analysis: The next generation of citation analysis. *Journal of the American Society for Information Science & Technology. 65*(9)

[6] Landauer, T.K., & Dumais, S.T., 1997. A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review. 104*(2). Pag 211–240

[7] Blei, D.M., Ng, A.Y., & Jordan, M.I., 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research. 3*(4-5). Pag 993–1022

[8] Miller, G.A., 1995. WordNet: A lexical database for English. *Communications of the ACM. 38*(11). Pag 39–41

[9] Budanitsky, A., & Hirst, G., 2006. Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics. 32*(1). Pag 13–47

[10] Dascalu, M., Dessus, P., Bianco, M., Trausan-Matu, S., & Nardy, A., 2014. Mining texts, learners productions and strategies with ReaderBench In A. Peña-Ayala (Ed.), *Educational Data Mining: Applications and Trends*, Springer. Switzerland. Pag 335–377

[11] Jurafsky, D., & Martin, J.H., 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*, Pearson Prentice Hall. London. 2nd ed.

[12] Porter, M., & Boulton, R. (2002). Snowball. http://snowball.tartarus.org/. Retrieved from http://snowball.tartarus.org/

[13] Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D., 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics. 39*(4)

[14] Dascalu, M., 2014. *Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence*, Springer. Switzerland

[15] Dascalu, M., Trausan-Matu, S., & Dessus, P., 2014. Validating the Automated Assessment of Participation and of Collaboration in Chat Conversations. In S. Trausan-Matu, K. E. Boyer, M. Crosby & K. Panourgia (Eds.), *12th Int. Conf. on Intelligent Tutoring Systems (ITS 2014)*. Springer. Pag 230–235

[16] Lupan, D., Dascalu, M., Trausan-Matu, S., & Dessus, P., 2012. Analyzing emotional states induced by news articles with Latent Semantic Analysis. In A. Ramsay & G. Agre (Eds.), *15th Int. Conf. on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2012)*. Springer. Pag 59–68

[17] O'Rourke, S.T., & Calvo, R.A., 2009. Analysing semantic flow in academic writing In V. Dimitrova, R. Mizoguchi, B. du Boulay & A. C. Graesser (Eds.), *Artificial Intelligence in Education. Building learning systems that care: From knowledge representation to affective modelling (AIED2009)*, IOS Press. Amsterdam, The Netherlands. Pag 173–180